Electronic Version

Stylesheet Version v1.1.1

DOCUMENT SIMILARITY DETECTION AND CLASSIFICATION SYSTEM

Abstract

A document similarity detection and classification system is presented. The system employs a case-based method of classifying electronically distributed documents in which content chunks of an unclassified document are compared to the sets of content chunks comprising each of a set of previously classified sample documents in order to determine a highest level of resemblance between an unclassified document and any of a set of previously classified documents. The sample documents have been manually reviewed and annotated to distinguish document classifications and to distinguish significant content chunks from insignificant content chunks. These annotations are used in the similarity comparison process. If a significant resemblance level exceeding a predetermined threshold is detected, the classification of the most significantly resembling sample document is assigned to the unclassified document. Sample documents may be acquired to build and maintain a repository of sample documents by detecting unclassified documents that are similar to other unclassified docu-

ments and subjecting at least some similar documents to a manual review and classification process. In a preferred embodiment the invention may be used to classify email messages in support of a message filtering or classification objective.